TITLE:

Exact Identification of the Structure of a Probabilistic Boolean Network from Samples

AUTHOR(S):

Cheng, Xiaoqing; Mori, Tomoya; Qiu, Yushan; Ching, Wai-Ki; Akutsu, Tatsuya

CITATION:

Cheng, Xiaoqing ...[et al]. Exact Identification of the Structure of a Probabilistic Boolean Network from Samples. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2016, 13(6): 1107-1116

ISSUE DATE:

2016-11-01

URL:

http://hdl.handle.net/2433/252326

RIGHT:

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.; この論文は出版社版でありません。引用の際には出版社版をご確認ご利用ください。; This is not the published version. Please cite only the published version.

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

Exact Identification of the Structure of a Probabilistic Boolean Network from Samples

Xiaoqing Cheng, Tomoya Mori, Yushan Qiu, Wai-Ki Ching, and Tatsuya Akutsu, *Member, IEEE*

Abstract—We study the number of samples required to uniquely determine the structure of a probabilistic Boolean network (PBN), where PBNs are probabilistic extensions of Boolean networks. We show via theoretical analysis and computational analysis that the structure of a PBN can be exactly identified with high probability from a relatively small number of samples for reasonably wide classes of PBNs of bounded indegree. On the other hand, we also show that there exist classes of PBNs for which it is impossible to uniquely determine the structure of a PBN from samples.

Index Terms—probabilistic Boolean networks, genetic networks, network inference, sample complexity

1 Introduction

Various kinds of mathematical models have been utilized for understanding dynamical behavior of biological systems. Among them, the *Boolean network* (BN) is a simple but well-studied model, which was proposed by Kauffman in 1969 as a model of gene regulatory networks [1], [2]. In a BN, each node takes a Boolean value, 0 or 1, at each time step, and the states of all nodes are updated synchronously according to Boolean functions assigned to nodes, where each node corresponds to a gene, and 1 and 0 mean that genes are active and inactive, respectively. In spite of the simplicity of the model, BNs exhibit complex behavior and thus extensive studies have been done in order to understand their behavior. For example, many studies have been done on the distribution of attractors [3], [4], [5] and the robustness against perturbations [6], [7].

In order to study realistic BNs, we need to infer the structures of BNs from real data such as gene expression time series data. Therefore, extensive studies have been done on inference of BNs from gene expression data [8], [9], [10], [11]. It is known that a BN with n nodes is uniquely determined with high probability from randomly selected $O(\log n)$ state-transition samples (i.e., $O(\log n)$ random arcs in the state transition diagram) if the maximum indegree (i.e., the maximum number of input nodes) is bounded by a constant [9], where $\log n$ stands for $\log_2 n$ in this paper. This is an interesting result because the possible number of BNs with n nodes is 2^{n2^n}, and 2^n samples are needed to uniquely specify a BN if there is no constraint on the structure of a BN (i.e., there is one-to-one correspondence between BNs and state transition diagrams).

Although BNs are a deterministic model, biological systems contain intrinsic stochasticity and observed data include noise. Therefore, BNs have been extended for including noise [13], [14], [15]. Among them, the *probabilistic Boolean network* (PBN) model has attracted much attention because of its simplicity, flexibility, and relations to Markov chains and Bayesian networks [16], [17]. In a PBN, multiple Boolean functions can be assigned to each node and one function is randomly selected at each time step according to the prescribed probability distribution. Extensive studies have been done on control and simulation of PBNs [18], [19], [20], [21], [22]. Several studies have also been done on inference of PBNs [16], [22], [23].

However, to our knowledge, there is no result on the sample complexity analogous to one for BNs. Although there exist some studies on related models (e.g., Bayesian networks) [14], [24], [25], results in [24], [25] are not on exact identification but on approximate identification, and the model in [14] is far from PBNs. Therefore, in this paper, we study the number of samples required to exactly identify the structure (i.e., a set of Boolean functions assigned to each node) of a PBN. We show that there are cases for which it is impossible to uniquely determine a PBN from samples. This result is reasonable because PBNs are stochastic systems. Interestingly, we also show that the structure of a PBN can be identified with high probability from $O(\log n)$ samples for reasonably wide classes of PBNs of bounded indegree.

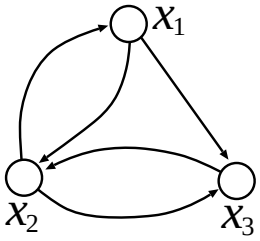
2 Probabilistic Boolean Network

A BN is a directed network with n nodes $x_1, \ldots, x_n$. Each node takes either 0 or 1 at each time step, and the state of x_i at time step t is denoted by $x_i(t)$. The states of all nodes are updated simultaneously according to Boolean functions assigned to nodes. Let $f^{(i)}$ denote the Boolean function assigned to node x_i, $IN(f^{(i)})$ denote a set of input nodes for $f^{(i)}$, and $\hat{f}^{(i)}$ denote its extension to all nodes $x_1, \ldots, x_n$. Let $\mathbf{x}(t)$ denote the global state of a BN at time t, (i.e., $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$. Then, the dynamics of the BN is given by $\mathbf{x}(t+1) = (\hat{f}^{(1)}(\mathbf{x}(t)), \ldots, \hat{f}^{(n)}(\mathbf{x}(t)))$. For example, suppose $IN(f^{(i)}) = \{x_j, x_k\}$. Then, $x_i(t+1)$ is determined by $x_i(t+1) = f^{(i)}(x_j(t), x_k(t))$, which can also be written as $x_i(t+1) = \hat{f}^{(i)}(\mathbf{x}(t))$.

- *X. Cheng, Y. Qiu, and W-K. Ching are with Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong, China. E-mail: mechxqxiaofen@gmail.com,{u3001131,wching}@hku.hk,*
- *T. Mori and T. Akutsu are with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. E-mail: {tmori,takutsu}@kuicr.kyoto-u.ac.jp.*

X.C. and T.M. contributed equally to this work.
Manuscript received XXX XX, XXXX; revised XXX XX, XXXX.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository



	Boolean Function	Prob.
$f_1^{(1)}$	$x_1(t+1) = x_2(t)$	0.8
$f_2^{(1)}$	$x_1(t+1) = \overline{x_2(t)}$	0.2
$f_1^{(2)}$	$x_2(t+1) = x_1(t) \wedge \overline{x_3(t)}$	0.7
$f_2^{(2)}$	$x_2(t+1) = x_3(t)$	0.3
$f_1^{(3)}$	$x_3(t+1) = x_1(t) \wedge \overline{x_2(t)}$	1.0

Fig. 1. Example of PBN.

A PBN is an extension of a BN. It is also a directed network with n nodes. As mentioned before, in a PBN, multiple Boolean functions can be assigned per node and one function is randomly selected at each time step according to the prescribed probability distribution. Let $\{f_1^{(i)}, \ldots, f_{m_i}^{(i)}\}$ be a set of Boolean functions assigned to node x_i, and $c_j^{(i)}$ denote the selection probability of $f_j^{(i)}$, where $\sum_{j=1}^{m_i} c_j^{(i)} = 1$ must be satisfied. It is to be noted that $IN(f_j^{(i)}) \neq IN(f_k^{(i)})$ is allowed. Then, $x_i(t+1)$ is determined by

$$x_i(t+1) = \hat{f}_j^{(i)}(\mathbf{x}(t)) \text{ with probability } c_j^{(i)},$$

where selection of $f_j^{(i)}$ is independent of selections in previous time steps and selections for other nodes. Since selections of Boolean functions are done for all nodes simultaneously and independently, dynamics of the whole PBN is represented by the transition probabilities from global states at time t to those at time $t+1$.

An example of PBN is given in Fig. 1. Suppose that $\mathbf{x}(t) = (0,0,0)$. If $(f_1^{(1)}, f_1^{(2)}, f_1^{(3)})$ is selected with probability $0.8 \times 0.7 = 0.56$, $\mathbf{x}(t+1) = (0,0,0)$ holds. Similarly, if $(f_1^{(1)}, f_2^{(2)}, f_1^{(3)})$ is selected with probability $0.8 \times 0.3 = 0.24$, $\mathbf{x}(t+1) = (0,0,0)$ holds. On the other hand, if $(f_2^{(1)}, f_1^{(2)}, f_1^{(3)})$ is selected with probability $0.2 \times 0.7 = 0.14$ or $(f_2^{(1)}, f_2^{(2)}, f_1^{(3)})$ is selected with probability $0.2 \times 0.3 = 0.06$, $\mathbf{x}(t+1) = (1,0,0)$ holds. Therefore, we have the following transition probabilities:

$$\begin{aligned} Prob(\mathbf{x}(t+1) = (0,0,0) \mid \mathbf{x}(t) = (0,0,0)) &= 0.8, \\ Prob(\mathbf{x}(t+1) = (1,0,0) \mid \mathbf{x}(t) = (0,0,0)) &= 0.2, \end{aligned}$$

where the probabilities of the other transitions from $(0,0,0)$ is 0.

3 Two Models

In this paper, we focus on PBNs in which two different Boolean functions are assigned to each node. Therefore, a PBN always means such one unless otherwise stated.

For a 0-1 bit vector $\mathbf{a}$, $\mathbf{a}_{+i}$ and $\mathbf{a}_{-i}$ denote the bit vector obtained from $\mathbf{a}$ by setting i-th bit of $\mathbf{a}$ to be 1 and 0, respectively, and $\mathbf{a}[i]$ denotes the value of i-th bit (i.e., 0-1 value corresponding to x_i). We say that x_i is *relevant* in f if there exists a 0-1 assignment $\mathbf{a}$ such that $\hat{f}(\mathbf{a}_{+i}) \neq \hat{f}(\mathbf{a}_{-i})$. When we consider a function f on K variables, we assume that all K-input variables are relevant in f, and K is called the *degree* of f ($K = |IN(f)|$).

We focus on one output node because pairs of Boolean functions can be identified independently for distinct output nodes. In the following, the target output node is denoted as y (i.e., $y = x_i$ for some i). A *sample* is defined by an assignment of 0-1 values to $x_1, \ldots, x_n$ and y, and is represented by a pair $(\mathbf{a}, b)$ of n-dimensional 0-1 input vector $\mathbf{a}$ and 0-1 (output) value b, where $\mathbf{a}$ corresponds to the global state at time t and b corresponds to the state of node y at time $t+1$.

We say that a Boolean function f and a sample $(\mathbf{a}, b)$ are *consistent* if $\hat{f}(\mathbf{a}) = b$. We also say that a pair of Boolean functions (f_1, f_2) and a sample are *consistent* if $\hat{f}_1(\mathbf{a}) = b$ or $\hat{f}_2(\mathbf{a}) = b$ holds. If a Boolean function or a pair of Boolean functions is consistent with every element in a set S of samples, it is also *consistent* with S. We say that (f_1, f_2) and S are *strongly consistent* if (f_1, f_2) and S are consistent and all possible consistent input/output pairs on $(IN(f_1) \cup IN(f_2), \{y\})$ appear in S. It is seen from the definitions that if (f_1, f_2) is strongly consistent with S, it is also consistent with S.

In the following, (f_1, f_2) is identified with (f_2, f_1), and a set of pairs of Boolean functions is referred as a *class of PBNs*. Furthermore, for simplicity, we assume that $\mathbf{a}$ in each sample is selected uniformly at random, and choice of f_1 and f_2 is also done uniformly at random. However, all the combinatorial results are independent from the probability distribution.

When analyzing the sample complexity, we assume that each sample is generated according to some underlying pair (f_1, f_2) of Boolean functions. Precisely, we assume that a *sample* is generated by a random 0-1 assignment to n variables and the corresponding value of f_1 or f_2. A *set of samples* is obtained by independent execution of this generation process where duplicate samples are unified.

We consider two models: *full information model* (FIM) and *partial information model* (PIM). In FIM, we say that a pair of Boolean functions (f_1, f_2) is *identified* from a set of samples S if (f_1, f_2) is the only one pair in the target class of Boolean functions that is strongly consistent with S. On the other hand, in PIM, we say that a pair of Boolean functions (f_1, f_2) is *identified* from a set of samples S if (f_1, f_2) is the only one pair in the target class of Boolean functions that is consistent with S. In either model, we say that a class C of PBNs is *identifiable* from samples if, for any pair (f_1, f_2) in C, there exists a sample set S such that (f_1, f_2) is identified from S.

Proposition 1. *If a class C of PBNs is identifiable from samples under PIM, C is also identifiable from samples under FIM.*

Proof. Let (f_1, f_2) be the correct pair. Since (f_1, f_2) can be identified under PIM, there exists a set S of samples for which only (f_1, f_2) is consistent with S under class C.

For any such S, there exists $S' \supseteq S$ such that (f_1, f_2) is strongly consistent with S'. Then, all other pairs are inconsistent with S' because they are inconsistent with S. □

This proposition implies that if C is not identifiable under FIM, it is not identifiable under PIM. It is straightforward to see that the following propositions hold.

Proposition 2. *A class C of PBNs is identifiable from samples under PIM if and only if for any two different pairs (f_1, f_2) and (f_3, f_4) in C there exists a 0-1 assignment $\mathbf{a}$ such that $|\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\}| > |\{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}|$ or $\hat{f}_1(\mathbf{a}) = \hat{f}_2(\mathbf{a}) \neq \hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a})$ holds.*

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

Proposition 3. *A class C of PBNs is identifiable from samples under FIM if and only if for any two different pairs (f_1, f_2) and (f_3, f_4) in C there exists a 0-1 assignment $\mathbf{a}$ such that $\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\} \neq \{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}$ holds.*

Here, we show classes of PBNs that are not identifiable from samples. Suppose that C is a set of pairs of one-input Boolean functions. That is, an element of C has the form of either (x_i, x_j), $(x_i, \overline{x_j})$, $(\overline{x_i}, \overline{x_j})$. Then, this class is not identifiable from samples under either model since $\{x_i, \overline{x_i}\} = \{x_j, \overline{x_j}\} = \{0, 1\}$ holds for any x_i, x_j. It should be noted that if samples are generated according to $(x_i, \overline{x_i})$, the value of y is determined as if it were completely at random and thus we cannot identify input node(s). This idea can be generalized to the class of pairs of AND functions (resp., OR functions) of degree 2.

Proposition 4. *The class of pairs of AND functions (resp., OR functions) of degree 2 is not identifiable from samples under PIM or FIM.*

Proof. Let $n = 3$. Consider two pairs of Boolean functions $(f_1, f_2) = (x_1 \wedge x_2, x_1 \wedge \overline{x_2})$ and $(f_3, f_4) = (x_1 \wedge x_3, x_1 \wedge \overline{x_3})$. Then, $\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\} = \{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}$ holds for all $\mathbf{a} \in \{0, 1\}^3$ as shown below, where "$0, 1$" means that we have both outputs.

x_1	0	0	0	0	1	1	1	1
x_2	0	0	1	1	0	0	1	1
x_3	0	1	0	1	0	1	0	1
f_1	0	0	0	0	0	0	1	1
f_2	0	0	0	0	1	1	0	0
Output	0	0	0	0	0, 1	0, 1	0, 1	0, 1
f_3	0	0	0	0	0	1	0	1
f_4	0	0	0	0	1	0	1	0

□

Next, we show another example to explain why we consider two models. Let $C = \{(x_1, x_1 \vee x_2), (x_1 \wedge x_2, x_1 \vee x_2)\}$. Consider the following two sample sets S_1 and S_2, where "$0, 1$" means that we have both outputs.

x_1	0	0	1	1
x_2	0	1	0	1
Output in S_1	0	0, 1	1	1
Output in S_2	0	0, 1	0, 1	1

Then, the former pair is strongly consistent with S_1, but is not consistent with S_2. The latter pair is strongly consistent with S_2 and is consistent with S_1. Therefore, the class C is identifiable from samples under FIM, but is not under PIM. However, when S_1 is given, only the former pair is strongly consistent. After a sample $(10, 0)$ is added (i.e., the set of sample becomes S_2), the former pair is no more strongly consistent but the latter pair newly becomes strongly consistent. Therefore, under FIM, we may not be able to know whether the current set of samples is enough. On the other hand, we can know under PIM whether the current set of samples is enough. However, as shown later, a much wider class of PBNs is identifiable under FIM. Therefore, it might be better to use FIM if a relatively large number of samples are available.

4 Partial Information Model

For PIM, we focus on the case in which each of f_1 and f_2 is an AND or OR function of fixed degree K. Based on Proposition 4, we assume that both x_i and $\overline{x_i}$ do not appear in f_1 and f_2 (i.e., if x_i appears in f_1, then $\overline{x_i}$ cannot appear in f_2) for any variable x_i. Such a pair is called an *admissible* AND/OR pair. We will show that the class of admissible AND/OR pairs of fixed degree K ($K > 1$) is identifiable from samples under PIM. From Proposition 2, it is enough to show that for any $(f_3, f_4) \neq (f_1, f_2)$, there exists $\mathbf{a}$ such that $|\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\}| > |\{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}|$ or $\hat{f}_1(\mathbf{a}) = \hat{f}_2(\mathbf{a}) \neq \hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a})$ holds. In the following, such an assignment is called a *witness* for $\langle (f_1, f_2), (f_3, f_4) \rangle$. It is to be noted that a witness for $\langle (f_1, f_2), (f_3, f_4) \rangle$ is not necessarily a witness for $\langle (f_3, f_4), (f_1, f_2) \rangle$.

We consider admissible AND/OR functions of degree K in Theorem 1. In the proof, we consider several cases depending on types (AND/OR) of f_is. The basic idea is common as follows. If there exists an assignment $\mathbf{a}$ such that $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0 \vee \hat{f}_2(\mathbf{a}) = 0$ (or, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ and $\hat{f}_1(\mathbf{a}) = 1 \vee \hat{f}_2(\mathbf{a}) = 1$) hold, $\mathbf{a}$ is a witness for $\langle (f_1, f_2), (f_3, f_4) \rangle$. To be more precise, we consider the case that all f_is are AND functions, $IN(f_1) = IN(f_3)$, and $IN(f_2) = IN(f_4)$. Since $(f_1, f_2) \neq (f_3, f_4)$, we can assume without loss of generality (w.l.o.g.) that there exists a variable x_i such that x_i appears negatively in f_1 but positively in f_3. Since (f_3, f_4) is an admissible pair, there exists an assignment $\mathbf{a}$ such that $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$. Then, $\hat{f}_1(\mathbf{a}) = 0$ holds for any such $\mathbf{a}$.

Theorem 1. *The class of PBNs of admissible AND/OR functions of degree K is identifiable from samples under PIM.*

Proof. We consider five cases depending on AND/OR types of f_i ($i = 1, \ldots, 4$), which cover all cases by taking symmetric cases (i.e., exchange of AND and OR) into account. For each case, we prove that there exists a witness for $\langle (f_1, f_2), (f_3, f_4) \rangle$.

Case 1: all f_is ($i = 1, \ldots, 4$) are AND, and $(f_3, f_4) \neq (f_1, f_2)$.

It is enough to consider the following two cases (see also Fig. 2).

Case 1-A: $IN(f_1) = IN(f_3)$ and $IN(f_2) = IN(f_4)$.
Since $(f_1, f_2) \neq (f_3, f_4)$, we can assume w.l.o.g. that there exists a variable x_i such that x_i appears negatively in f_1 but positively in f_3. Since (f_3, f_4) is an admissible pair, there exists an assignment $\mathbf{a}$ such that $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$. Then, $\hat{f}_1(\mathbf{a}) = 0$ holds for any such $\mathbf{a}$, and thus $\mathbf{a}$ is a witness.

Case 2-A: $IN(f_1) \neq IN(f_3)$ and $IN(f_1) \neq IN(f_4)$.
We assume w.l.o.g. from the symmetry and the discussion in Case 1-A that all variables appear positively in f_1, f_2, f_3, and f_4. Let $x_i \in IN(f_3) \backslash IN(f_1)$ and $x_j \in IN(f_4) \backslash IN(f_1)$, where $x_i = x_j$ is allowed, and $X_1 \setminus X_2$ denotes the set $\{x | x \in X_1, x \notin X_2\}$. Consider an assignment $\mathbf{a}$ such that 0 is assigned to x_i and x_j, and 1 is assigned to the other variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ and $\hat{f}_1(\mathbf{a}) = 1$ hold.

Case 2: f_1, f_2, f_3 are AND functions, and f_4 is an OR function.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

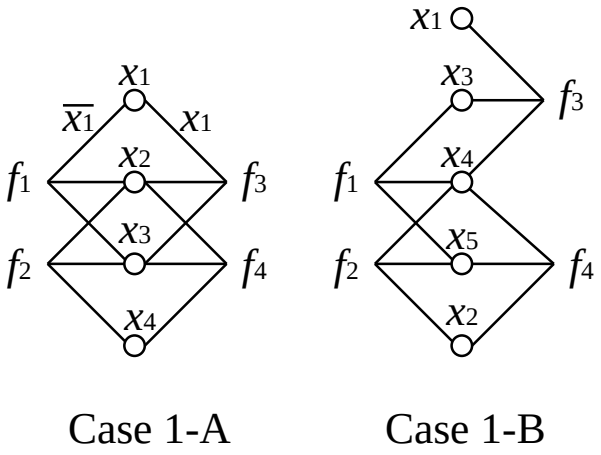


Fig. 2. Illustration of two subcases of Case 1 in the proof of Theorem 1. In Case 1-A, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ but $\hat{f}_1(\mathbf{a}) = 0$ hold for an assignment $\mathbf{a} = (1, 1, 1, 1)$. In Case 1-B, $f_3 = f_4 = 0$ but $f_1 = 1$ for an assignment $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ but $\hat{f}_1(\mathbf{a}) = 1$ hold for an assignment $\mathbf{a} = (0, 0, 1, 1, 1)$.

We assume w.l.o.g. that $IN(f_2) \neq IN(f_3)$. We further assume w.l.o.g. that all variables appear positively in f_1, f_2, f_3, and f_4.

Let $x_i \in IN(f_2) \setminus IN(f_3)$. Consider an assignment $\mathbf{a}$ such that 0 is assigned to x_i and 1 is assigned to the other variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_2(\mathbf{a}) = 0$ hold.

Case 3: f_1, f_3, f_4 are AND functions, and f_2 is an OR function.

We assume w.l.o.g. that $IN(f_2) \neq IN(f_3)$. We further assume w.l.o.g. that all variables appear positively in f_1, f_2, f_3, and f_4.

Let $x_i \in IN(f_2) \setminus IN(f_3)$. Consider an assignment $\mathbf{a}$ such that 1 is assigned to x_i and 0 is assigned to the other variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ and $\hat{f}_2(\mathbf{a}) = 1$ hold.

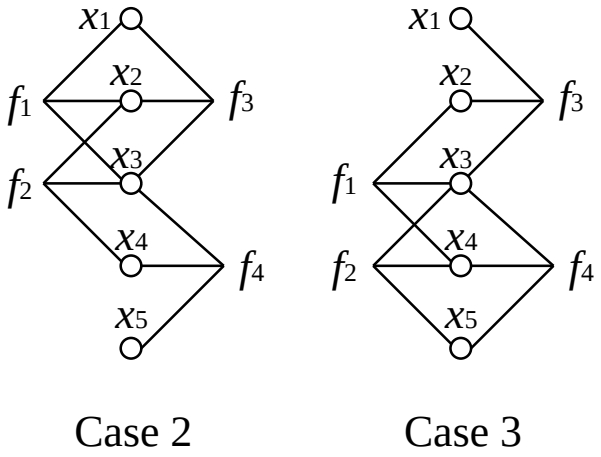


Fig. 3. Illustration of Case 2 and Case 3 in the proof of Theorem 1. In Case 2, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_2(\mathbf{a}) = 0$ hold for an assignment $\mathbf{a} = (1, 1, 1, 0, 1)$. In Case 3, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ and $\hat{f}_2(\mathbf{a}) = 1$ hold for an assignment $\mathbf{a} = (0, 0, 0, 0, 1)$.

Case 4: f_1, f_2 are AND functions, and f_3, f_4 is an OR function.

We assume w.l.o.g. that all variables appear positively.

Choose an arbitrary variable $x_i \in IN(f_1)$. Consider an assignment $\mathbf{a}$ such that 0 is assigned to x_i and 1 is assigned to the other variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ hold.

Case 5: f_1, f_3 are AND functions, and f_2, f_4 is an OR function.

It is enough to consider the following cases.

Case 5-A: $IN(f_1) = IN(f_3)$ and $IN(f_2) = IN(f_4)$.
Since $(f_1, f_2) \neq (f_3, f_4)$, we assume w.l.o.g. that there exists x_i such that x_i appears negatively in f_1 or f_2 and all variables in $IN(f_3) \cup IN(f_4)$ appear positively in f_3 and f_4. If x_i appears negatively in f_1 (resp., in f_2), consider an assignment $\mathbf{a}$ such that 1 (resp., 0) is assigned to all variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ (resp., $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ and $\hat{f}_2(\mathbf{a}) = 1$) hold.

Case 5-B: $IN(f_1) \neq IN(f_3)$ (resp., $IN(f_2) \neq IN(f_4)$).
We consider only the case of $IN(f_1) \neq IN(f_3)$, where the other case can be proved in an analogous manner. We assume w.l.o.g. that all variables appear positively in each f_i.

Let $x_i \in IN(f_1) \setminus IN(f_3)$. Consider an assignment $\mathbf{a}$ such that 0 is assigned to x_i and 1 is assigned to the other variables. Then, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ hold.

□

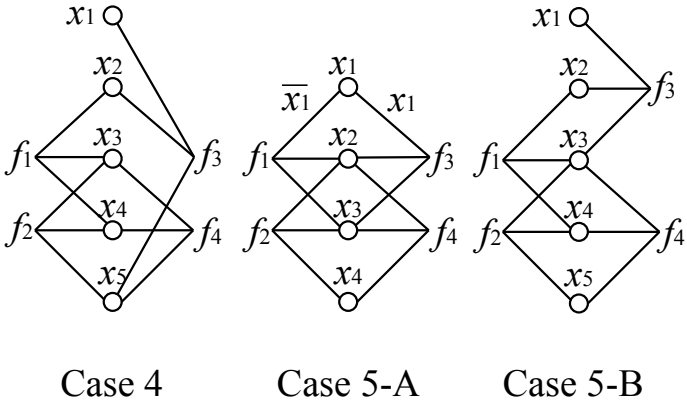


Fig. 4. Illustration of Case 4 and Case 5 in the proof of Theorem 1. In Case 4, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ hold for an assignment $\mathbf{a} = (1, 0, 1, 1, 1)$. In Case 5-A, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ hold for an assignment $\mathbf{a} = (1, 1, 1, 1)$. In Case 5-B, $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ and $\hat{f}_1(\mathbf{a}) = 0$ hold for $\mathbf{a} = (1, 1, 1, 0, 1)$.

The theorem cannot be generalized to cases in which a pair of Boolean functions of different degrees is assigned. Consider the case of $f_1 = x_1 \wedge x_2 \wedge x_3$, $f_2 = x_2 \wedge x_3 \wedge x_4$, $f_3 = x_1 \wedge x_2$, and $f_4 = x_2 \wedge x_3 \wedge x_4$. Then, any assignment $\mathbf{a}$ with $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 1$ makes $\hat{f}_1(\mathbf{a}) = \hat{f}_2(\mathbf{a}) = 1$, and any assignment $\mathbf{a}$ with $\hat{f}_3(\mathbf{a}) = \hat{f}_4(\mathbf{a}) = 0$ makes $\hat{f}_1(\mathbf{a}) = \hat{f}_2(\mathbf{a}) = 0$. It is also impossible to generalize the theorem to cases in which a different number of Boolean functions can be assigned. Consider the case of (f_1, f_2) and (f_3, f_4, f_5)

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

with $f_1 = f_3 = x_1 \wedge x_2$, $f_2 = f_4 = x_2 \wedge x_3$, $f_5 = x_3 \wedge x_4$. Since $f_1 = f_3$ and $f_2 = f_4$, samples consistent with (f_1, f_2) are always consistent with (f_3, f_4, f_5).

5 Full Information Model

For FIM, we consider the class C_K of pairs of Boolean functions (f_1, f_2) such that $f_1 = \ell_s \vee f_1'$ and $f_2 = \ell_t \wedge f_2'$ where ℓ_r is either x_r or $\overline{x_r}$, $x_s \neq x_t$, and f_i' $(i = 1, 2)$ is any Boolean function such that $\{x_s, x_t\} \cap IN(f_i') = \{\}$ and $2 \leq |IN(f_i')| \leq K - 1$. This class is called the class of *complementary canalyzing pairs* (of degree K). Recall that a Boolean function f is called a canalyzing function if f can be represented as either $f = x_i \vee f'$ or $f = x_i \wedge f'$, where it is suggested that most biologically meaningful Boolean functions are canalyzing ones [26]. Then, we have the following theorem.

Theorem 2. *For any positive integer $K \geq 3$, C_K is not identifiable from samples under PIM but is identifiable from samples under FIM.*

Proof. First, we consider PIM. Let $f_1 = x_1 \vee (x_2 \wedge x_3)$, $f_2 = x_4 \wedge (x_2 \vee x_3)$, $f_3 = x_1 \vee x_2 \vee x_3$, and $f_4 = x_4 \wedge x_2 \wedge x_3$. Then, it is straight-forward to verify that (f_3, f_4) is consistent with any sample generated from (f_1, f_2). Therefore, C_3 (or, C_K with $K > 3$) is not identifiable from samples under PIM.

Next, we consider FIM. Let $f_1 = x_s \vee f_1'$, $f_2 = x_t \wedge f_2'$. The proof can be easily modified for the cases that x_s and/or x_t appear negatively. In the following $f_i = f_j$ means that $\hat{f}_i(\mathbf{a}) = \hat{f}_j(\mathbf{a})$ holds for all assignments $\mathbf{a}$. It is to be noted that f_i and f_j can have different representations.

Let $f_3 = \ell_p \vee f_3'$ and $f_4 = \ell_q \wedge f_4'$, where ℓ_p (resp., ℓ_q) is either x_p or $\overline{x_p}$ (resp., x_q or $\overline{x_q}$). Then, it is enough to show that if $(f_1, f_2) \neq (f_3, f_4)$, there exists an assignment $\mathbf{a}$ such that $\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\} \neq \{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}$. We consider the following cases, where symmetric cases are omitted.

Case 1: $x_s = x_p = \ell_p$.

Suppose that $f_2 \neq f_4$. Since $x_s = x_p$, $x_s \notin IN(f_2)$ and $x_s \notin IN(f_4)$ hold. Then, there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$ and $\hat{f}_2(\mathbf{a}) \neq \hat{f}_4(\mathbf{a})$. We assume w.l.o.g. that $\hat{f}_2(\mathbf{a}) = 0$ and $\hat{f}_4(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$.

Suppose that $f_2 = f_4$. Then, $f_1 \neq f_3$ holds and thus there exists an assignment $\mathbf{a}$ such that $\hat{f}_1(\mathbf{a}) \neq \hat{f}_3(\mathbf{a})$. Since $\hat{f}_2(\mathbf{a}) = \hat{f}_4(\mathbf{a})$ holds, we have $\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\} \neq \{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}$.

Case 2: $x_s = x_p$ and $\ell_p = \overline{x_p}$.

Consider an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 0$, $\mathbf{a}[t] = 0$, and $\hat{f}_1'(\mathbf{a}) = 0$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (0, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$.

Case 3: $x_s = x_q = \ell_q$.

Suppose that there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 0$, $\mathbf{a}[t] = 0$, $\mathbf{a}[p] = 1$ and $\hat{f}_1'(\mathbf{a}) = 0$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (0, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$. Therefore, we can assume w.l.o.g. that $f_1' = x_p \vee f_1''$ and $f_1 = x_s \vee x_p \vee f_1''$.

Next, suppose that there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$, $\mathbf{a}[t] = 0$, $\mathbf{a}[p] = 1$ and $\hat{f}_4'(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$. Therefore, we can assume w.l.o.g. that $f_4' = x_t \wedge f_4''$ and $f_4 = x_s \wedge x_t \wedge f_4''$.

Assuming $f_1 = x_s \vee x_p \vee f_1''$, suppose that there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 0$, $\mathbf{a}[t] = 1$, $\mathbf{a}[p] = 1$ and $\hat{f}_2'(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 1)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$. Therefore, we can assume w.l.o.g. that $f_2' = \overline{x_p} \wedge f_2''$ and $f_2 = x_t \wedge \overline{x_p} \wedge f_2''$.

Here, suppose that there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$, $\mathbf{a}[t] = 1$, $\mathbf{a}[p] = 1$ and $\hat{f}_4''(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$.

If $\hat{f}_4''(\mathbf{a}) = 0$ holds for all $\mathbf{a}$, we can consider an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$, $\mathbf{a}[t] = 1$, and $\hat{f}_2'(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 1)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (0, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$.

Case 4: $x_s = x_q$ and $\ell_q = \overline{x_q}$.

Consider an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$, $\mathbf{a}[t] = 1$, and $\hat{f}_2'(\mathbf{a}) = 1$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 1)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (0, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$.

Case 5: $x_s \neq x_p$, $x_s \neq x_q$, $x_t \neq x_p$, $x_t \neq x_q$.

We consider the case of $x_s = \ell_s$, $x_t = \ell_t$, $x_p = \ell_p$, and $x_q = \ell_q$, where the other cases can be proven in a similar way.

If $x_p \notin IN(f_1')$, we can consider an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 0$, $\mathbf{a}[t] = 0$, $\mathbf{a}[p] = 1$, and $\hat{f}_1'(\mathbf{a}) = 0$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (0, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$.

In the following, we assume w.l.o.g. that $x_p \in IN(f_1')$. Suppose that there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 0$, $\mathbf{a}[t] = 0$, $\mathbf{a}[p] = 1$, and $\hat{f}_1'(\mathbf{a}) = 0$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (0, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$. Therefore, we can assume that f_1 has the form of $f_1 = x_s \vee x_p \vee f_1''$.

Suppose that $x_s \notin IN(f_3')$. We can consider an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = 1$, $\mathbf{a}[p] = 0$, $\mathbf{a}[q] = 0$, and $\hat{f}_3'(\mathbf{a}) = 0$. Then, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 0)$ or $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (1, 1)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (0, 0)$. Therefore, we assume w.l.o.g. that $x_s \in IN(f_3')$ holds. Then, f_3 has the form of $f_3 = x_s \vee x_p \vee f_3''$ as in the case of f_1.

If $f_1'' = f_3''$ (i.e., $f_1 = f_3$), we can prove the theorem as in Case 1. Therefore, we assume w.l.o.g. that $f_1'' \neq f_3''$. Then, there exists an assignment $\mathbf{a}$ such that $\mathbf{a}[s] = \mathbf{a}[p] = 0$, $\hat{f}_1''(\mathbf{a}) = 0$, and $\hat{f}_3''(\mathbf{a}) = 1$. If $\mathbf{a}[t] = 0$, we have $(\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})) = (0, 0)$ whereas $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 0)$ or $(\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})) = (1, 1)$. If there exists no such assignment, f_3'' has the form of $f_3'' = x_t \wedge f_3'''$ and thus $f_3 = (x_s \vee x_p) \vee (x_t \wedge f_3''')$. Similarly, we can assume w.l.o.g. that $f_1 = (x_s \vee x_p) \vee (x_q \wedge f_1''')$, $f_2 = (x_t \wedge x_q) \wedge (x_p \vee f_2''')$, and $f_4 = (x_t \wedge x_q) \wedge (x_s \vee f_4''')$. Then, there exists an assignment $\mathbf{a}$ such that either $\hat{f}_1'''(\mathbf{a}) = 1$, $\hat{f}_2'''(\mathbf{a}) = 0$, $\hat{f}_3'''(\mathbf{a}) = 1$, or $\hat{f}_4'''(\mathbf{a}) = 0$ holds, otherwise $f_1 = f_3$ and $f_2 = f_4$ would hold. We can assume w.l.o.g. that $\hat{f}_1'''(\mathbf{a}) = 1$ holds. Consider the assignment $\mathbf{a}' = \mathbf{a}_{-s,-p,-t,+q}$. Then, we have $(\hat{f}_1(\mathbf{a}'), \hat{f}_2(\mathbf{a}')) = (1, 0)$ whereas $(\hat{f}_3(\mathbf{a}'), \hat{f}_4(\mathbf{a}')) = (0, 0)$. □

We provide an example showing that Theorem 2 cannot be generalized to the class of PBNs of nested canalyzing

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

TABLE 1
Difficult case for nested canalyzing functions under FIM.

x_1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
x_2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
x_3	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
x_4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
f_1	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1
f_2	0	0	1	1	0	1	1	1	0	0	1	1	0	1	1	1
Output	0	0	0,1	0,1	0	0,1	1	1	0	0	0,1	0,1	0,1	1	1	1
f_3	0	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1
f_4	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1

function pairs consisting of only positive literals. Let $K = 3, n = 4$, consider two pairs of Boolean functions $(f_1, f_2) = (x_2 \wedge (x_1 \vee x_3), x_3 \vee (x_2 \wedge x_4))$ and $(f_3, f_4) = (x_3 \vee (x_1 \wedge x_2), x_2 \wedge (x_4 \vee x_3))$. Then, $\{\hat{f}_1(\mathbf{a}), \hat{f}_2(\mathbf{a})\} = \{\hat{f}_3(\mathbf{a}), \hat{f}_4(\mathbf{a})\}$ holds for all $\mathbf{a} \in \{0,1\}^4$ as shown in Table 1, where "0, 1" means that we have both outputs.

6 Sample Complexity

In the above, we have shown that the classes of admissible AND/OR pairs and complementary canalyzing pairs can be identified under PIM and FIM, respectively Here, we analyze how many samples are required in order to uniquely identify a PBN under the following assumptions:

- a sample is a pair $(\mathbf{a}, \mathbf{b})$ of 0-1 assignment to n variables and 0-1 output values of n nodes,
- each 0-1 assignment is generated uniformly at random,
- the output value of each node is determined according to a pair of Boolean functions assigned to the node in an underlying PBN of the target class, where selection of a function from the pair is independently done for each assignment with probability $1/2$,
- each sample is generated independently.

Note that $\mathbf{a}$ and $\mathbf{b}$ in a sample $(\mathbf{a}, \mathbf{b})$ correspond to global states of a PBN at time steps t and $t+1$, respectively, although samples are independent from each other. Using Theorem 1, Theorem 2, and a similar argument as in [9], we have the following theorem.

Theorem 3. *For the class of PBNs consisting of admissible AND/OR pairs of degree K (resp., complementary canalyzing pairs of degree K), if $O(2^{4K+1}(4K+\alpha)\log n)$ samples are given uniformly at random, the correct PBN can be uniquely identified with probability at least $1 - \frac{1}{n^\alpha}$ under PIM (resp., FIM).*

Proof. We prove for the case of PIM, where the case of FIM can be proved in an analogous manner.

Suppose that (f_1, f_2) is the underlying function pair for k-th node in a PBN. It is seen from the proof of Theorem 1 that if all possible assignments on $I = IN(f_1) \cup IN(f_2) \cup IN(f_3) \cup IN(f_4)$ and their possible output values by f_1 and f_2 are given, inconsistency of (f_3, f_4) $(\neq (f_1, f_2))$ can be detected. Since $|I| \leq 4K$, it is also seen that if all possible assignments on all possible combination of $4K$ variables and their possible output values appear in samples (Condition C1), we can uniquely identify (f_1, f_2).

The probability that $\mathbf{a}[i] = 1$ for $i = 1, \ldots, 4K$ and $\mathbf{b}[k] = \hat{f}_1(\mathbf{a})$ do not hold in a given sample $(\mathbf{a}, \mathbf{b})$ is at most $1 - \frac{1}{2^{4K+1}}$, and thus the same condition does not hold in any m samples is at most $(1 - \frac{1}{2^{4K+1}})^m$. Since the number of combination of $4K$ variables is less than n^{4K}, the probability that Condition C1 does not hold is bounded above by $2^{4K+1} \cdot n^{4K} \cdot (1 - \frac{1}{2^{4K+1}})^m$. Since there exist n nodes, the probability that Condition C1 does not hold for one or more nodes is bounded above by

$$p_{K,n,m} = 2^{4K+1} \cdot n^{4K+1} \cdot \left(1 - \frac{1}{2^{4K+1}}\right)^m.$$

It is not difficult to see that $p_{K,n,m} \leq p$ holds if

$$m > \ln 2 \cdot 2^{4K+1}((4K+1)(1+\log n) + \log \frac{1}{p}).$$

Letting $p = \frac{1}{n^\alpha}$, the theorem holds.

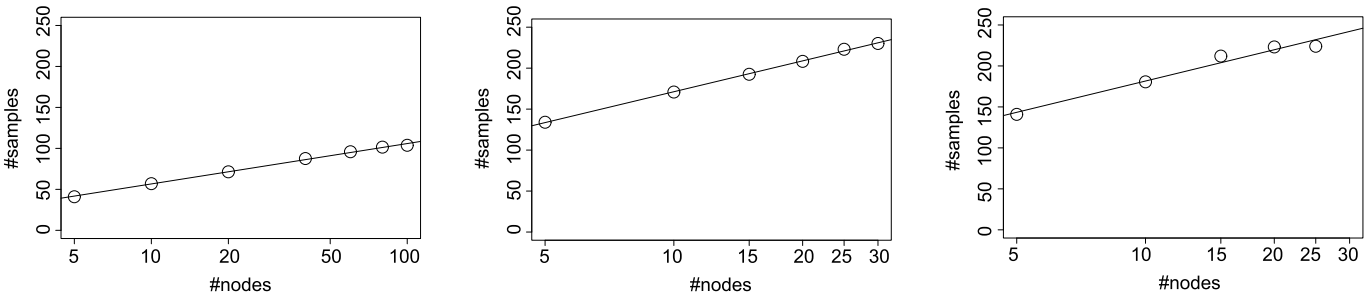
□

7 Computational Experiments

In order to verify the result of Theorem 3, we performed computational experiments. Since it is impossible under FIM to know whether a given sample set is enough, we only examined the case of PIM. In the experiments, we examined the cases of $K = 1$ and $K = 2$, where both types of PBNs consisting of AND functions (AND PBNs) and AND/OR functions (AND/OR PBNs) were tested for $K = 2$. For each n and K, we randomly generated 100 PBNs. Then, for each PBN, we generated samples by assigning 0 or 1 to each node with probability $1/2$ and updating the PBN synchronously to obtain output values. Finally, we computed the average number of samples required to uniquely identify each PBN. For that purpose, we examined samples one-by-one until the number of consistent PBNs became 1, where we maintained consistent Boolean function pairs independently for each node. The programs for generation and identification of PBNs were implemented using Python language, whereas simulation of PBNs was performed using 'BoolNet' [27]. All experiments were performed on a PC cluster with Intel(R) Xeon(R) CPU E5-2690 2.90GHz and 35.87 GB memory.

The results are shown in Fig. 5. For the case of $K = 1$ (Fig. 5(a)), n is varied from 5 to 100. It seems that the number of samples is proportional to $\log n$. Even for the case of $n = 100$, approximately only 100 samples were required to identify PBNs. Although the computation time

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository



(a) $K = 1$ (b) $K = 2$ (AND PBNs) (c) $K = 2$ (AND/OR PBNs)

Fig. 5. Results of computational experiments. (a), (b) and (c) show the number of samples required to identify PBNs of the target class under PIM for the cases of $K = 1$, $K = 2$ (AND PBNs) and $K = 2$ (AND/OR PBNs), respectively. Regression lines are indicated within each panel, where x-axis is log-scaled.

for identification increased rapidly as n grew, it took around only 30 seconds per identification of a PBN.

Similarly, for the cases of $K = 2$ (AND PBNs) (Fig. 5(b)) and $K = 2$ (AND/OR PBNs) (Fig. 5(c)), the number of required samples also looks to be proportional to $\log n$. However, it took long computation time: approximately 12 hours and 36 hours per identification of a PBN for the largest cases of AND PBNs ($n = 30$) and AND/OR PBNs ($n = 25$), respectively. For this reason, we could not perform computational experiments for much larger n and K. It is also to be noted that the numbers of required samples for AND PBNs and AND/OR PBNs were not so different. This is reasonable because the set of all possible Boolean functions per node for an AND/OR PBN is at most three times larger than that for an AND PBNs.

8 Conclusion

We have studied the number of samples in order to exactly identify the structure of a PBN from samples, with focusing on the case where two Boolean functions are assigned to each node. We considered two models: partial information model (PIM) and full information model (FIM). We showed that it is possible to identify the structure of a PBN from a small number ($O(\log n)$) of samples under both models for reasonably wide classes of PBNs whereas there exist classes of PBNs whose structure cannot be identified from samples. In addition, the positive result on PIM was verified via computational experiments. These results are interesting because they show that the structure of a probabilistic system can be exactly identified to some extent. It is to be noted that these results are independent of inference algorithms. It was also shown that there is a large gap of identifiable classes between PIM and FIM. The merit of PIM is that we can know when the structure is uniquely identified. However, FIM allows us to identify the structure for a much wider class of PBNs. Therefore, if an enough number of samples are available, it is better to use FIM.

Although we obtained fundamental results on the sample complexity for identification of a PBN, there are many things to be explored. One important thing is to study the cases in which more than two Boolean functions are assigned per node. Another important thing is to make use of the probabilities assigned to Boolean functions. In order to show negative results, we assumed that the probability assigned to each function is $1/2$. However, if different probabilities are assigned, we may identify the structure of a PBN for wider classes of PBNs. In addition to the sample complexity issue, it is also important to develop efficient identification algorithms because we employed a naive enumeration-based algorithm in computational experiments and thus could not handle large-scale PBNs even for $K = 2$. Further studies might lead to much deeper understanding of complex biological networks.

Acknowledgments

This work was partially supported by Grant-in-Aid #26540125 from JSPS, Japan. We thank Avraham Melkman for helpful discussions.

References

[1] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, pp. 437-467, 1969.
[2] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution.* New York:Oxford University Press, 1993.
[3] B. Samuelsson and C. Troein, "Superpolynomial growth in the number of attractors in Kauffman networks," *Phys. Rev. Lett.*, vol. 90, 098701, 2003.
[4] B. Drossel, T. Mihaljev, and F. Greil, "Number and length of attractors in a critical Kauffman model with connectivity one," *Phys. Rev. Lett.*, vol. 94, 088701, 2005.
[5] P. Krawitz and I. Shmulevich, "Basin entropy in Boolean network ensembles," *Phys. Rev. Lett.*, vol. 98, 158701, 2007.
[6] C. Seshadhri, Y. Vorobeychik, J. R. Mayo, R. C. Armstrong, and J. R. Ruthruff, "Influence and dynamic behavior in random Boolean networks," *Phys. Rev. Lett.*, vol. 107, 108701, 2011.
[7] S. Squires, E. Ott, and M. Girvan, "Dynamical instability in Boolean networks as a percolation problem," *Phys. Rev. Lett.*, vol. 109, 085701, 2012.
[8] S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures," *in Proceedings of Pacific Symposium on Biocomputing 1998*, pp. 18–29, Singapore:World Scientific, 1998.
[9] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," *in Proceedings of Pacific Symposium on Biocomputing 1999*, Singapore:World Scientific, pp. 17-28. 1999.
[10] R. Laubenbacher and B. Stigler, "A computational algebra approach to the reverse engineering of gene regulatory networks," *J. Theor. Biol.*, vol. 229, pp. 523-537, 2004.
[11] T. J. Perkins and M. T. Hallett, "A trade-off between sample complexity and computational complexity in learning Boolean networks from time-series data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 7, pp. 118-128, 2010.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

[12] P. Vera-Licona1, A. S. Jarrah, L. D. Garcia-Puente, J. McGee, and R, Laubenbacher, "An algebra-based method for inferring gene regulatory networks." *BMC Syst. Biol.*, vol. 8, 37, 2014.
[13] E. N. Miranda and N. Parga, "Noise effects in the Kauffman model," *Europhys. Lett.*, vol. 10, pp. 293-298, 1989.
[14] T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinformatics*, vol. 16, pp. 727-734, 2000.
[15] T. P. Peixoto, "Redundancy and error resilience in Boolean networks," *Phys. Rev. Lett.*, vol. 104, 048701, 2010.
[16] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, pp. 261-274, 2001.
[17] I. Shmulevich and E. R. Dougherty, *Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks.* Philadelphia:SIAM, 2010.
[18] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, pp. 169-191, 2003.
[19] W-K. Ching, S. Zhang, M. Ng, and T. Akutsu, "An approximation method for solving the steady-state probability distribution of probabilistic Boolean networks," *Bioinformatics*, vol. 23, pp. 1511-1518, 2007.
[20] B. Faryabi, G. Vahedi, A. Datta, J. F. Chamberland, and E. R. Dougherty, "*Curr. Genomics*, vol. 10, pp. 463-477, 2009.
[21] K. Kobayashi and K. Hiraishi, "An integer programming approach to optimal control problems in context-sensitive probabilistic Boolean networks," *Automatica*, vol. 47, pp. 1260-1264, 2011.
[22] P. Trairatphisan, A. Mizera, J. Pang, A. A. Tantar, J. Schneider, and T. Sauter, "Recent development and biomedical applications of probabilistic Boolean networks," *Cell Commun. Signal.*, vol. 11, 46, 2013.
[23] S. Marshall, L. Yu, Y. Xiao, and E. R. Dougherty, "Inference of a probabilistic Boolean network from a single observed temporal sequence," *EURASUP J. Bioinform. Syst. Biol.*, vol. 2007, 32454, 2007.
[24] N. Friedman and Z. Yakhini, "On the sample complexity of learning Bayesian networks," *in Proceedings of 20th Conf. Uncertainty in Artificial Intelligence* pp. 274–282, CA:Morgan Kaufmann, 1996.
[25] S. Dasgupta, "The sample complexity of learning fixed-structure Bayesian networks," *Mach. Learn.*, vol. 29, pp. 165-180, 1997.
[26] S. E. Harris, B. K. Sawhill, A. Wuensche, and S. Kauffman, "A model of transcriptional regulatory networks based on biases in the observed regulation rules," *Complexity*, vol. 7, pp. 23-40, 2002.
[27] C. Müssel, M. Hopfensitz, and H. Kestler, "Boolnet–an R package for generation, reconstruction and analysis of boolean networks," *Bioinformatics*, vol. 26, pp. 1378-1380, 2010.

Xiaoqing Cheng was born in Hangzhou, China in 1990. She received the B.E degree from Department of Mathematics, University of Science and Technology of China, Hefei, China in 2012 and has been a Ph.D student in Department of Mathematics, The University of Hong Kong from 2012.09. Her main areas of research interest are bioinformatics and data mining.

Tomoya Mori received B.E. and M.E. degrees in informatics from Doshisha University and Kyoto University in 2010 and 2012, respectively. He is currently a doctor course student at Kyoto University, Japan. His research interests are in bioinformatics and discrete algorithms.

Yushan Qiu was born in Guangdong, China, in 1988. She received the B.E. degree in School of Mathematics from the South China Normal University, Guangdong, China, in 2007. She has been a PhD student in the Department of Mathematics, The University of Hong Kong since 2011. Her research interests include Bioinformatics and machine learning.

Wai-Ki Ching Wai-Ki Ching is a Professor and the head of department in the Department of Mathematics, University of Hong Kong. He obtained his B. Sci. (Hons) and M. Phil. degrees from the University of Hong Kong. He then received his Ph.D. degree from the Chinese University of Hong Kong. His research interests are stochastic modelling and matrix computations. In particular, the applications of mathematical models and numerical algorithms in solving problems related to Markov chains, bioinformatics, systems biology and management science.

Tatsuya Akutsu received the B.E. and M.E. degrees in aeronautics and the D.E. degree in information engineering from the University of Tokyo, 1984, 1986, and 1989, respectively. Since 2001, he has been a professor in the Bioinformatics Center, Institute for Chemical Research, Kyoto University. His research interests include bioinformatics and discrete algorithms. His research interests include bioinformatics and discrete algorithms. He is a member of the IEEE.